From Biophysics to Evolutionary Genetics: Statistical Aspects of Gene Regulation

Michael Lässig

Institut für Theoretische Physik, Universität zu Köln,
Zülpicher Str. 77, 50937 Köln, Germany
lassig@thp.uni-koeln.de

1 Introduction

Genomic functions often cannot be understood at the level of single genes but require the study of gene networks. This systems biology credo is nearly commonplace by now. Evidence comes from the comparative analysis of entire genomes: Current estimates put, for example, the number of human genes at around around 22000, hardly more than the 14000 of the fruit fly, and not even an order of magnitude higher than the 6000 of baker's yeast. The complexity and diversity of higher animals therefore cannot be explained in terms of their gene numbers. If, however, a biological function requires the concerted action of several genes, and conversely, a gene takes part in several functional contexts, an organism may be defined less by its individual genes but by their interactions. The emerging picture of the genome as a strongly interacting system with many degrees of freedom brings new challenges for experiment and theory, many of which are of a statistical nature. And indeed, this picture continues to make the subject attractive to a growing number of statistical physicists.

Genes encode proteins, and proteins perform functions in the cell. Hence, a gene takes part in a biological function only if it is *expressed*, i.e., if the protein produced from it is present in the cell. Genes interact by *regulation*: the protein of one gene can influence the production of protein from another gene. Gene regulation can take place during *transcription*, the process by which the cell reads the information contained in a gene and copies it to messenger RNA (which is subsequently used to make a functional protein). This is the most fundamental level of interactions between genes: the transcription of one gene may be enhanced or reduced by the expression of other genes. Transcriptional regulation is thus a good starting point for theory. We should keep in mind, however, that it is not the only mode of gene interactions. Especially in eukaryotes, additional regulation mechanims involving histones, chromatin, micro-RNAs etc. become relevant, which are just entering the stage of model building. An excellent introduction to the biology of regulation can be found in [1].

This article is a primer on theoretical aspects of gene interactions, and we limit ourselves to transcriptional regulation. Clearly, the subject has rather diverse aspects:

(1) Transcription is a *biophysical* process, which involves the interaction of DNA and proteins. Its regulation takes place through the binding of proteins to DNA at specific loci in the vicinity of the gene to be regulated. Already at this level, this process is rather complex and not yet fully understood. What enables the protein to find one or a few specific functional sites in a genome of up to billions of base pairs, bind there with sufficient strength to influence transcription, and leave again once its task is performed?

(2) Given the protein can find its functional sites, can we as well? If that is possible, we can predict the specific gene interactions building regulatory networks from sequence data. The analysis of regulatory DNA is a major topic of research in *bioinformatics*, with the aim of identifying statistical characteristics of functional loci and of building search algorithms.

(3) Regulation is also becoming an important part of *evolutionary biology* [2, 3]. If regulatory networks are to explain the differentiation of higher animals, there must be efficient modes of evolution for the interactions between genes. At the level of regulary DNA, these modes remain largely to be explored. It is clear, however, that the underlying evolutionary dynamics is the basis of a quantitative understanding of regulatory networks.

All three aspects of regulation contribute to a unified theoretical picture. Key concepts such as the biophysical binding energy, the bioinformatic scoring function, and the evolutionary fitness turn out to be rather deeply related. We will focus on these crosslinks between different fields, which are likely to become important for future research. A challenge for an introductory presentation is the diversity of relevant background material, only a rather ecclectic account of which can be presented here. Yet, I hope it transpires even from this short introduction that present quantitative genomics is an area of science shaped by a remarkable confluence of ideas from different disciplines.

2 Biophysics of transcriptional regulation

The fundamental step in the regulatory interaction between two genes is a binding process: the protein produced by the first gene acts as a *transcription factor* for the second gene, i.e., it binds to a functional site on the DNA close to the second gene and thereby enhances or suppresses its transcription. Binding sites are short, typically segments of 10 to 15 base pairs in prokaryotes and even shorter segments in eukaryotes. They are primarily located in the *cis-regulatory region* of a gene, which lies just upstream of its protein-coding sequence and extends over hundreds of base pairs in prokaryotes and over thousands of base pairs in eukaryotes. The scenario of transcriptional regulation is sketched in Fig. 1. A transcription factor bound to a functional binding site regulates the downstream gene by recruiting or repelling RNA polymerase. This protein-protein interaction catalyzes or suppresses the process of transcription of the gene. All these binding processes should not be

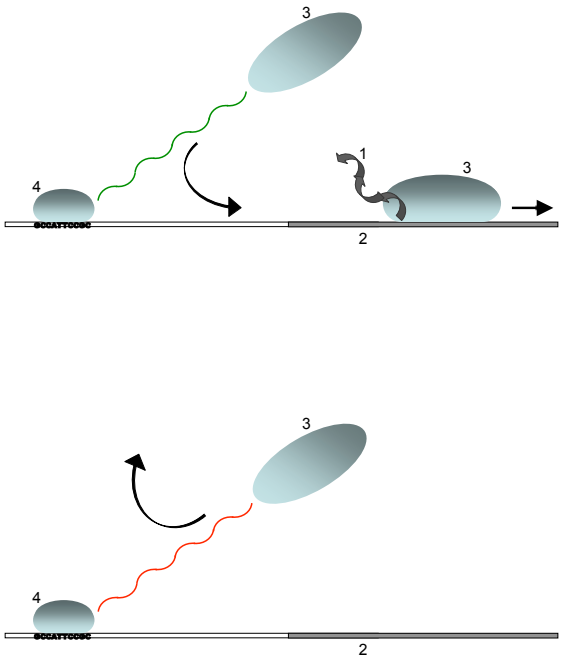

Fig. 1. Transcriptional regulation. Transcription is the synthesis of messenger RNA (1) whose genetic code is a copy of the coding DNA (2) of a gene, by means of RNA polymerase (3). A transcription factor (4) bound to a DNA target site interacts with RNA polymerase molecules, (a) enhancing or (b) reducing the transcription rate of a nearby gene.

understood as on or off; they happen with certain probabilities, which are determined by the binding energies and the numbers of the molecules involved.

Factor-DNA binding energies. The interaction of a transcription factor protein with DNA is two-fold: There is a position-unspecific attraction with energy E_u and a specific interaction, whose energy depends on the particular locus where the factor binds. The unspecific part is the electrostatic interaction between the positively charged protein and the negatively charged DNA backbone, while the specific part involves hydrogen bonds between the binding domain of the protein and the nucleotides of the binding locus. A locus is specified by its starting position r and its length ℓ (with relevant values ℓ of order 10). The specific binding energy $E(r)$ depends on ℓ consecutive nucleotides $\mathbf{a} = (a_1, \ldots, a_\ell)$ counted downstream from the starting position, the *sequence state* or *genotype* of that locus. Switching between unspecific and

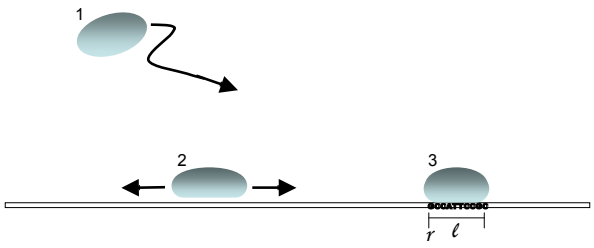

Fig. 2. Thermodynamic states of a transcription factor. (1) Unbound state, with three-dimensional diffusion. (2) Unspecific bound state, with one-dimensional diffusion along the DNA backbone. (3) Specific bound state. The binding energy depends on the genotype at the binding locus, which has length ℓ and whose position is specified by the coordinate r.

specific binding takes place via a conformation change of the factor protein. As a result of these interactions, the factor protein can be in three thermodynamic states as shown in fig. 2: unbound (i.e., freely diffusing), unspecifically bound (i.e., diffusing along the DNA backbone), and specifically bound.

The biophysics of factor-DNA binding has been established in a series of seminal papers [4, 5, 6, 7]. More recently, the characteristics of specific binding have been measured for some bacterial transcription factors [8, 9, 10, 11, 12]. These can be summarized as follows:

(a) The single nucleotides of a binding locus $\mathbf{a} \equiv (a_1, \ldots, a_\ell)$ give approximately independent contributions to the binding energy,

$$E(\mathbf{a}) = \sum_{i=1}^{\ell} \epsilon_i(a_i). \tag{1}$$

(b) At each position i, there is typically one preferred nucleotide a_i^* with $\epsilon_i(a_i^*) = \min_a \epsilon_i(a)$. Hence, there is a unique "ground state" sequence $\mathbf{a}^* = (a_1^*, \ldots a_\ell^*)$ with minimal binding energy $E^* \equiv E(\mathbf{a}^*)$, i.e., with strongest binding.

(c) Mismatches with respect to the minimum-energy sequence involve energy costs $\epsilon_i(a) - \epsilon_i(a_i^*) \approx 1 - 3\ k_B T$ per nucleotide.

(d) There is an energy difference $E_u - E^* \sim 15\ k_B T$ between unspecific and strongest specific binding.

Experimental data for the binding energies $\epsilon_i(a)$ are known only for a few transcription factors. Approximate values for these energies can also be inferred from nucleotide frequencies in functional binding sites [10]. For order-of-magnitude estimates, one often uses the so-called two-state approximation [7], which is homogeneous in the nucleotide positions and distinguishes

only between match and mismatch:

$$\epsilon_i(a) - \epsilon_i(a_i^*) = \begin{cases} \epsilon & \text{if } a_i \neq a_i^* \\ 0 & \text{if } a_i = a_i^* \end{cases} \tag{2}$$

with $\epsilon \approx 2k_BT$. In this approximation, the binding energy of a sequence $\mathbf{a}$ is simply related to the *Hamming distance* $d(\mathbf{a}, \mathbf{a}^*)$, i.e., the number of nucleotide mismatches between $\mathbf{a}$ and $\mathbf{a}^*$,

$$E(\mathbf{a}) = E^* + \epsilon \cdot d(\mathbf{a}, \mathbf{a}^*). \tag{3}$$

Energy distribution in the genome. Fig. 3(a) shows the sequence of energy values $E(r)$ found in a segment of the *E. coli* genome for a specific transcription factor, the cAMP response protein (CRP) This "energy landscape" looks quite random, i.e., consecutive energy values are approximately uncorrelated. The distribution $W_{\text{dat}}(E)$ of energies over the entire noncoding part of the *E. coli* genome is shown in fig. 3(b). We can compare this with the distribution $W_0(E)$ obtained from a random sequence with the same nucleotide frequencies (i.e., from a scrambled genome). The distribution $W_0(E)$ is approximately Gaussian as expected for a sum of independent random variables ϵ_i according to eq. (1). The actual distribution $W_{\text{dat}}(E)$ is indeed of the same form as $W_0(E)$ for most energies. However, a closer look at the low-energy tail of the distribution shows that there are significantly more strong binding sites than expected from a random sequence [13, 14, 15]. So at least some of them are there not by chance but for a reason.

Search kinetics. All three thermodynamic modes of a factor molecule - free diffusion, unspecific binding, and specific binding - are important for the search kinetics towards a functional site [4, 5, 6]. The unspecific attraction causes the transcription factor to be bound to DNA with a finite probability, i.e., a given molecule spends about equal amounts of time on and off the DNA backbone. Hence, the search process is a mixture of effectively one-dimensional diffusion along the DNA backbone and three-dimensional diffusion in the surrounding medium. This proves more efficient than purely one- or three-dimensional diffusion. In the 1D mode, the factor diffuses in a flat energy landscape if it is in the conformation of unspecific binding, or in the landscape $E(r)$ if it is in the conformation of specific binding. In this way, it can sample the low-energy part of the landscape $E(r)$ while avoiding its barriers. The main obstacles on its way to a functional site are spurious binding sites, which have a low energy $E(r)$ by chance and act as traps. We lack a completely satisfactory picture of the search kinetics, which is an area of current research [13, 16]. However, this process proves to be remarkably fast. Typical search times are less than a minute, i.e., substantially shorter than typical functional intervals in a cell cycle of at least minutes. Therefore, the regulatory effect of a site is related to its probability of binding a factor molecule at equilibrium, which can be evaluated by standard thermodynamics.

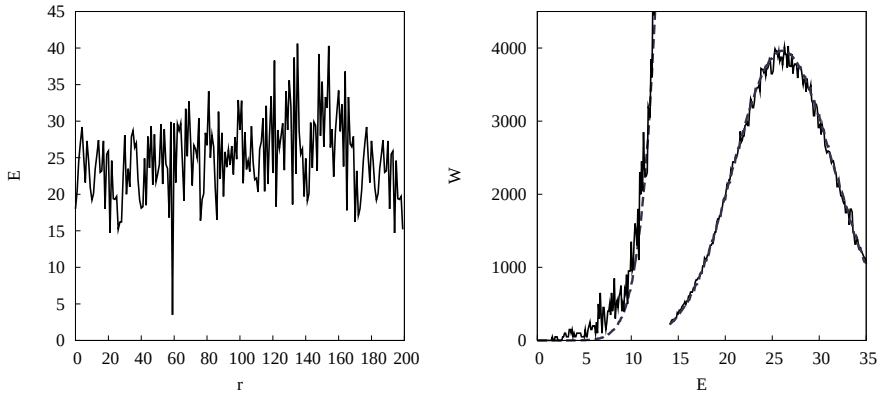

Fig. 3. Transcription factor binding energies of the *E. coli* genome. (a) Energy "landscape" $E(r)$ for specific binding of the CRP factor at 200 consecutive positions r in an intergenic region, with a binding site at position 59. (b) Count histogram $W_{\rm dat}(E)$ with energy bins of width 0.1 obtained from all intergenic regions, together with the distribution $W_0(E)$ for a random sequence (dashed line, shown with a 30fold zoom into the region $E < 14$). From [15].

Thermodynamics of factor binding. We start with the idealized but instructive problem of a single factor protein interacting with a genome of length $L \gg 1$, which contains a single functional site, while the rest of the sequence is random. Since the protein is bound to the DNA with a probability of about 1/2, we neglect the unbound state for the subsequent probability estimates and study only the bound protein, which is at equilibrium between specific and unspecific binding. At each position r, the likelihood of these two states is given by the Boltzmann factors $\exp[-E(r)/k_BT]$ and $\exp[-E_u/k_BT]$, respectively. Hence, the partition function for a single protein has the form

$$Z = \sum_{r=1}^{L} \mathrm{e}^{-E(r)/k_BT} + L\,\mathrm{e}^{-E_u/k_BT}. \tag{4}$$

The functional site, which is assumed to be positioned at $r = r_f$, must have a low specific binding energy $E \equiv E(r_f)$. We now single out this position and write

$$\begin{aligned} Z &= \mathrm{e}^{-E/k_BT} + \sum_{r\neq r_f} \mathrm{e}^{-E(r)/k_BT} + L\,\mathrm{e}^{-E_u/k_BT} \\ &\approx \mathrm{e}^{-E/k_BT} + Z_0, \end{aligned} \tag{5}$$

where Z_0 is the partition function of a completely random sequence. The probability of the factor being bound specifically at the functional site is then

$$p(E) = \frac{\mathrm{e}^{-E/k_BT}}{Z} = \frac{1}{1+\mathrm{e}^{(E-F_0)/k_BT}}, \tag{6}$$

where $F_0 = -k_BT \log Z_0$ is the free energy for a random genome. Thus, the binding probability depends on the binding energy in a sigmoid way, with a threshold energy $E = F_0$ between strong and weak binding. This strongly nonlinear dependence is known to physicists as a Fermi function.

It is easy to generalize the thermodynamic formalism to more than one factor molecule. Ignoring the overlap between close sites, each position r can be empty or be occupied either by an unspecifically or by a specifically bound factor. Using a chemical potential σ, the many-factor partition function can hence be written as

$$Z(\sigma) = \prod_{r=1}^{L} Z(\sigma, r), \tag{7}$$

where $Z(\sigma, r)$ is a sum over the three thermodynamic states at position r,

$$Z(\sigma, r) = 1 + \mathrm{e}^{\sigma - E(r)/k_BT} + \mathrm{e}^{\sigma - E_u/k_BT}. \tag{8}$$

The chemical potential σ is determined by the number of factor molecules, n, via the relation $n = (\mathrm{d}/\mathrm{d}\sigma) \log Z(\sigma)$. For actual transcription factor numbers, which are of order $1 - 10^4$, this relation is well approximated by [13]

$$\sigma = \frac{F_0}{k_BT} + \log n. \tag{9}$$

The functional site is now occupied by a specifically bound factor with probability

$$p(E) = \frac{\mathrm{e}^{\sigma - E/k_BT}}{Z(\sigma, r_f)} = \frac{1}{1+\mathrm{e}^{(E-F_0)/k_BT - \log n}}. \tag{10}$$

The binding probability - and hence the effects of the functional site on the regulated gene - are thus determined by the binding energy, the number of factor molecules, and on the genomic background (via the free energy F_0). The dependence $p(E)$ is a Fermi function with threshold energy $E = F_0 + k_BT \log n$, which is shifted with respect to the single-molecule case. Clearly, p is also a Fermi function of $\log n$ at fixed binding energy, with a threshold at $\log n = (E - F_0)/k_BT$.

Sensitivity and genomic design of regulation. The regulatory machinery can be very efficient: in bacteria, it has been shown that single factor molecules can have regulatory effects. We can use eq. (6) to enquire how the cell can reach this high level of sensitivity, following mostly ref. [13]. We assume a minimal genome which has a single functional site of maximum binding strength E^* and is otherwise random. If a single factor molecule is to affect regulation, its binding to the functional site must not be overwhelmed by the remainder of the genome. This leads to a criterion on the signal-to-noise ratio of regulatory interactions,

$$F_0 \gtrsim E^*, \tag{11}$$

which in turn imposes a number of constraints on the design of regulatory DNA:

(a) In a random genome, there must be at most of order one minimum-energy binding sites, i.e., $L(1/4)^\ell \gtrsim 1$. This gives a lower bound on the site length, $\ell \gtrsim \log L / \log 4$. For a bacterial genome ($L \sim 10^6$), we obtain $\ell \gtrsim 10$, which gives the right length of functional binding sites. However, this bound is not fulfilled in eukaryotes. Indeed, eukaryotic genomes use a different design with groups of adjacent binding sites.

(b) For each minimum-energy site, there are ℓ suboptimal sites of Hamming distance 1 from the minimum-energy sequence. These must not suppress the binding to the minimum-energy site, i.e., $\exp(-E^*/k_BT) \gtrsim \ell \exp[-(E^* + \epsilon)/k_BT]$ in the two-state approximation. This gives a lower bound on the binding energy per nucleotide, $\epsilon/k_BT \gtrsim \log \ell \approx 2 - 3$.

(c) Finally, the unspecific binding in the entire genome must not suppress the specific binding to a minimum-energy site, i.e., $\exp(-E^*/k_BT) \gtrsim L \exp(-E_u/k_BT)$. This produces a lower bound on the energy gap between unspecific and optimal specific binding, $(E_u - E^*)/k_BT \gtrsim \log L \approx 15$.

Quite remarkably, these bounds are fulfilled as approximate equalities in bacteria. Hence, the machinery of transcriptional regulation operates just at the treshold of single-molecule sensitivity, i.e, $F_0 \approx E^*$.

Programmability and evolvability of regulatory networks. Of course, not every regulatory interaction is equally sensitive. To switch genes on or off, the cell uses the dependencies of the binding probability both on factor numbers and on binding energies. During the cell cycle, the level of n can vary over several orders of magnitude, say, between a few and tens of thousands of molecules. At a given value of n, the effects on the regulated genes differ since their functional sites have different values of E. The binding energies can change on evolutionary time scales by mutations of the site sequence, which leads to regulatory differences between individuals and, ultimately, between species. Both parameters are thus necessary to encode pathways in regulatory networks. This is most flexible if minimum-energy sites are indeed sensitive to a single factor molecule as discussed above. Differential *programmability* as a network design principle [13] thus favors complicated molecular structures with longer binding sites and larger binding energies. However, this competes with the *evolvability* of the system by a stochastic evolution process [17]. We have seen that the single-molecule sensitivity is just marginally reached in bacteria. This indicates that the actual machinery may result from a compromise between programmability and evolvability: binding sites are just complicated enough to work. It also indicates that genomic structures can only be understood from their evolution; this aspect will be developed further in Section 4.

3 Bioinformatics of regulatory DNA

Predicting regulatory interactions between genes is clearly a key problem in bioinformatics, which is as important as the analysis of individual genes and proteins. It is not surprising that this problem is very difficult since, as we have discussed in the last section, targeting regulatory input in a large genome is a tremendous signal-to-noise problem even for the cell itself. Its solution via the analysis of regulatory DNA requires finding statistical criteria to distinguish between functional binding sites and background sequence. A general introduction to the relevant sequence statistics can be found in ref. [18].

Markov model for background sequence. We begin by specifying a stochastic model for the nonfunctional segments of intergenic DNA. These are assumed to be Markov sequences with uniform single-nucleotide frequencies $p_0(a)$ $(a = A, C, G, T)$. Hence, the probability of finding a given sequence has the factorized form

$$P_0(a_1, \ldots, a_k) = \prod_{i=1}^{k} p_0(a_i). \tag{12}$$

This assumption should not be taken too literally. The term "nonfunctional" refers to binding of a particular transcription factor. Intergenic DNA contains plenty of non-random elements with other functions (e.g., binding sites for other factors) or without known function (such as repeat elements). The salient point is, however, that most of intergenic DNA is well approximated by a Markov sequence with respect to binding of a given transcription factor. To make this more precise, we project the distribution $P_0(\mathbf{a})$ for segments of length ℓ onto the binding energy E as independent variable. Denoting the projected distribution for simplicity with the same letter P_0, we have

$$P_0(E) \equiv \sum_{\mathbf{a}} P_0(\mathbf{a})\, \delta(E - E(\mathbf{a})). \tag{13}$$

This distribution is close to the actual genomic distribution $W_{\mathrm{dat}}(E)$ for most values of E, as we have seen in fig. 3. It is possible to improve the background model by introducing small frequency couplings between neigboring letters [14, 15].

Probabilistic model for functional sites. The sequences $\mathbf{a} = (a_1, \ldots, a_\ell)$ at functional sites of a given transcription factor are assumed to be drawn from a different distribution $Q(\mathbf{a})$. We write this distribution in the form

$$Q(\mathbf{a}) = P_0(\mathbf{a}) \exp[S(\mathbf{a})]. \tag{14}$$

The quantity $S(\mathbf{a})$, which is called the *relative log likelihood score* of the distributions P_0 and Q, will turn out to have an important evolutionary meaning as well.

The single-nucleotide distribution $q_i(a)$ at a given position i within functional loci is obtained by summing the full distribution Q over all other positions

$$q_i(a) = \sum_{a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_\ell} Q(\mathbf{a}). \tag{15}$$

The set of these marginal distributions, $q_i(a)$ $(i = 1, \ldots, \ell;\ a = A, C, G, T)$ is called the *position weight matrix* for binding sites of a given factor [19]. If the score function is additive in the nucleotide positions, $S(\mathbf{a}) = \sum_{i=1}^{\ell} s_i(a_i)$, the Q distribution has a factorized form, $Q(\mathbf{a}) = \prod_{i=1}^{\ell} q_i(a_i)$ with

$$q_i(a) = p_0(a) \exp[s_i(a)]. \tag{16}$$

This additivity assumption is made in most of the existing literature since the position weight matrix (15) can be inferred from a sample of known functional site sequences, which in turn determines directly the single nucleotide scores (16). This scoring is the basis for a number of site prediction methods in single species and by cross-species analysis; see, e.g., refs. [19, 20, 21, 22, 23].

Here we treat functional sites as coherent statistical units and do not make the assumption of additivity of the score function [15]. As will be discussed in the next section, functionality imposes correlations between the nucleotide frequencies within a functional site, preventing factorization of the Q distribution. Of course, it is not possible to reconstruct the full distribution $Q(\mathbf{a})$, which lives on a 4^ℓ-dimensional sequence space, from a limited sample of experimentally known functional sites. However, we can again project this distribution onto the binding energy as independent variable, $Q(E) \equiv \sum_{\mathbf{a}} Q(\mathbf{a})\delta(E - E(\mathbf{a}))$. Since all regulatory effects of a functional site depend on its sequence a only via the binding energy, we can also write the score as a function of the energy, $S(\mathbf{a}) = S(E(\mathbf{a}))$ (this will become obvious in the next section). Hence, the relationship (14) has the same form for the projected distributions,

$$Q(E) = P_0(E) \exp[S(E)]. \tag{17}$$

Bayesian model for genomic loci. Assuming that functional loci are distributed randomly with a small probability λ, we now combine the models for background sequence and for functional sites into a model for the full distribution of sequences $\mathbf{a}$ in intergenic DNA,

$$W(\mathbf{a}) = (1 - \lambda)P_0(\mathbf{a}) + \lambda Q(\mathbf{a}). \tag{18}$$

(At the moment, we are ignoring the possible overlap between functional sites). In the language of statistics, this is a probabilistic model with *hidden variables*. The output of this model consists of pairs $(m, \mathbf{a})$: First, the model variable $m \in \{f, 0\}$ is drawn with probabilities λ and $1 - \lambda$ (i.e., a locus is labelled as nonfunctional or functional), then the sequence is drawn from

the corresponding distribution $P_0(\mathbf{a})$ or $Q(\mathbf{a})$. However, only the sequence counts $\mathbf{a}$ are available data. The "hidden" variable m can be inferred from the data in a probabilistic way using Bayes' formula, which expresses the joint probability distribution of data and model in terms of its conditional and its marginal distributions

$$\text{prob}(\mathbf{a}, m) = \text{prob}(\mathbf{a}|m)\,\text{prob}(m) = \text{prob}(m|\mathbf{a})\,\text{prob}(\mathbf{a}) \tag{19}$$

with $\text{prob}(\mathbf{a}) = \sum_m \text{prob}(\mathbf{a}|m)\text{prob}(m)$. We can solve for the conditional probability of the model for given data $\mathbf{a}$,

$$\text{prob}(m|\mathbf{a}) = \frac{\text{prob}(\mathbf{a}|m)\,\text{prob}(m)}{\sum_m \text{prob}(\mathbf{a}|m)\,\text{prob}(m)}. \tag{20}$$

For the probability of functionality, $\rho_f(\mathbf{a}) \equiv \text{prob}(f|\mathbf{a})$, this formula reads

$$\rho_f(\mathbf{a}) = \frac{\lambda Q(\mathbf{a})}{W(\mathbf{a})} = \frac{1}{1 + \exp[-S(\mathbf{a}) + \log\frac{1-\lambda}{\lambda}]}. \tag{21}$$

The dependence on S has again the form of a Fermi function. Its threshold value $S = \log[(1-\lambda)/\lambda]$ separates sequences that are more likely to be functional or more likely to be background.

The full Bayesian model (18) can again be projected onto the energy variable,

$$W(E) = (1-\lambda)P_0(E) + \lambda Q(E). \tag{22}$$

In this form, it can be tested against genomic data [15]. To plot the distributions P_0, Q, and W as functions of E, we use eq. (1) with an energy matrix $\epsilon_i(a) = \epsilon_0 \log[q_i(a)/p_0(a)]$ estimated from the position weight matrix up to an overall constant ϵ_0 [10]. For our example of the CRP transcription factor, the distribution $Q(E)$ can be estimated from the about 50 known binding sites in the *E. coli* genome. Using this Q distribution and a probability of functionality $\lambda \approx 6 \times 10^{-4}$, the full distribution $W(E)$ produces an excellent fit of the count histogram $W_{\text{dat}}(E)$ over the entire range of energies; see fig. 4(a). The log likelihood score function $S(E) = \log[Q(E)/P_0(E)]$ is shown in fig. 4(b), shifted such that the curve has its zero at a point $E_s \approx 13$ beyond which binding becomes negligible.

The resulting probability of functionality $\rho_f(E)$ as given by eq. (21) is also shown in fig. 4(b). This indicates the dilemma for the prediction of individual binding sites based on sequence data from a single species. Many functional sites have energies in the "twilight" region between the ensembles λQ and $(1-\lambda)P_0$, where ρ_f takes values around $1/2$. Hence, depending on the energy cutoff chosen, any prediction is torn between many false negatives or many false positives.

Dynamic programming and sequence analysis. It is straightforward to generalize the Bayesian approach to longer segments of intergenic DNA,

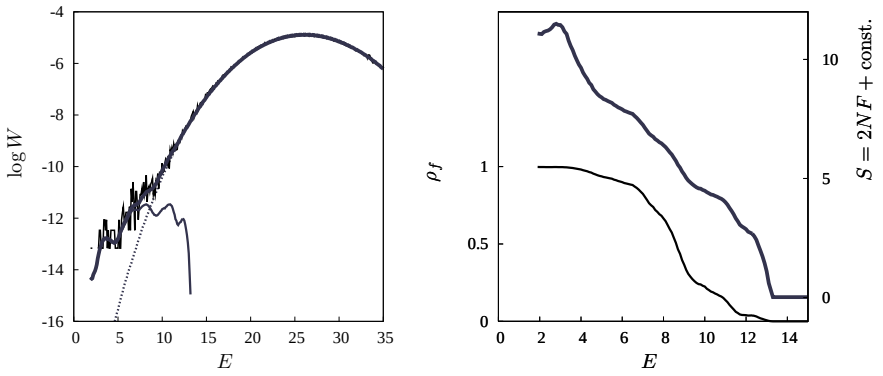

Fig. 4. Bayesian model for regulatory DNA and score function. (a) Energy count histogram $W_{\rm dat}(E)$ for CRP sites in *E. coli* as in fig. 3 (log scale), model distribution $W(E)$ (thick line), and its decomposition (22) into background component $(1-\lambda)P_0(E)$ (thin dashed line) and component $\lambda Q(E)$ ($E < E_s \approx 13$) of functional sites (thin solid line). (b) Log-likelihood score $S(E) = \log[Q(E)/P_0(E)]$ (shifted by a constant, thick line) and probability of functionality $\rho_f(E)$ (thin line). From [15].

which are covered by an unknown number s of non-overlapping functional sites as shown in fig. 5 [21]. The hidden variables are now the sequence of left initial positions $\mathbf{r}_f \equiv (r_1, \ldots, r_s)$ of the functional sites (with the no-overlap constraint $r_{\nu+1} \geq r_\nu + \ell$ for $\nu = 1, \ldots, s-1$). The full sequence distribution in a segment of length L has the form

$$W_L(a_1, \ldots, a_L) = Z^{-1} \sum_{\mathbf{r}_f} \tilde{\lambda}^s W_L(a_1, \ldots, a_L | \mathbf{r}_f), \tag{23}$$

where Z is a normalization factor, $\tilde{\lambda} = \lambda + O(\lambda^2)$ is a weight factor for each functional locus (the negligible correction terms originate from the no-overlap constraint), and $W_L(a_1, \ldots, a_L|\mathbf{r}_f)$ is the sequence distribution for given positions of functional loci,

$$\begin{aligned} &W_L(a_1, \ldots, a_L|\mathbf{r}_f) = \\ &\quad p_0(a_1) \ldots p_0(a_{r_1-1}) \prod_{\nu=1}^{s} Q(a_{r_\nu}, \ldots, a_{r_\nu+\ell-1})\, p_0(a_{r_\nu+\ell}) \ldots p_0(a_{r_{\nu+1}-1}) = \\ &\quad p_0(a_1) \ldots p_0(a_L) \exp\left[\sum_{\nu=1}^{s} S(a_{r_\nu}, \ldots, a_{r_\nu+\ell-1})\right] \end{aligned} \tag{24}$$

with $r_{n+1} \equiv L+1$. The sum over sequences $\mathbf{r}_f$ of arbitrary length s seems formidable at first, but W_L is easy to compute from the recursion

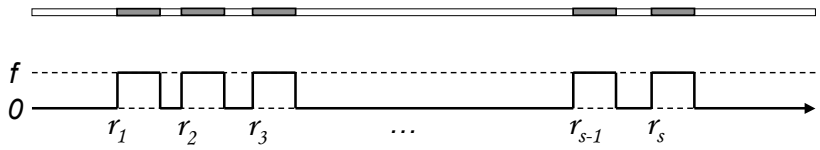

Fig. 5. Analysis of regulatory sequences. A configuration of s nonoverlapping binding sites is given by the sequence of left initial positions $\mathbf{r}_f = (r_1, \ldots, r_s)$ (with $r_{\nu+1} - r_\nu \geq \ell$ for $\nu = 1, 2, \ldots, s-1$). It can be associated with a path $m(r)$ which takes the values $m = f$ at the nucleotide positions of binding sites and $m = 0$ elsewhere. Dynamic programming algorithms based on a Bayesian model (24) of genomic sequences assign to each site configuration a probability of occurence $\rho(\mathbf{r}|a_1, \ldots, a_L)$ for given sequence data $a_1, \ldots, a_L$; see eq. (26).

$$W_r(a_1, \ldots, a_r) = (1 - \hat{\lambda}) p_0(a_r) W_{r-1}(a_1, \ldots, a_{r-1}) + \tilde{\lambda} Q(a_{r-\ell+1}, \ldots, a_r) W_{r-\ell}(a_1, \ldots, a_{r-\ell}) \qquad (25)$$

with the initial condition $W_0 = 1$ and $\hat{\lambda} = \tilde{\lambda} + O(\tilde{\lambda}^2)$. This type of recursion relation is usually called a *dynamic programming algorithm* in computer science. In physics, it is known as a *transfer matrix*, and the sum (24) is recognized as the corresponding discrete path integral in imaginary time r, if we interpret $\mathbf{r}_f$ as encoding a path $m(r)$ that takes the value $m = f$ at the nucleotide positions $r_\nu, \ldots, r_\nu + \ell - 1$ ($\nu = 1, \ldots, s$) within functional loci and $m = 0$ otherwise (see fig. 5). Both concepts prove very useful also in more general problems of sequence alignment.

In analogy to (21), the probability of a set $\mathbf{r}_f$ of functional loci for given sequence data is

$$\rho(\mathbf{r}_f|a_1, \ldots, a_L) = \frac{W_L(a_1, \ldots, a_L|\mathbf{r}_f)}{W_L(a_1, \ldots, a_L)}. \qquad (26)$$

The most likely set $\mathbf{r}_f^*$ can be obtained by the following "backward" algorithm: Given the sequence $(W_1, \ldots, W_L)$ obtained from the "forward" recursion (25), we can decide for every point r whether it is more likely to be a background position or the endpoint of a functional locus, ignoring all sequence information from positions $> r$. This depends on whether the leading contribution to W_r comes from the first or second term on the r.h.s. of (25) and defines the local optimum model $m^*(r)$. The global optimum set of functional loci respecting the no-overlap constraint is then $\mathbf{r}_f^* = \{r|b(r) = 1\}$, where $b(r)$ is given by the recursion $b(r) = \ell$ if $b(r+1) \leq 1$ & $m^*(r) = f$ and $b(r) = \max(b(r+1) - 1, 0)$ otherwise, with the initial condition $b(L+1) = 0$.

The Bayesian model can easily be extended to sequences containing several types of binding sites, which bind different transcription factors and are distinguished by their Q distributions. Dynamic programming algorithms can

thus predict the likely coverage of a sequence with binding sites of known type [21]. This is the first step in extending the statistical analysis from single binding sites to entire regions of regulatory DNA. Indeed, models of this kind have been applied successfully to predict regulatory elements in eukaryotes, which typically consist of functional groups of adjacent binding sites. In the algorithms currently used, however, the scoring in (24) is strictly additive for groups of non-overlapping binding sites: it does not take into account dependencies between the sites within one functional group or overlapping sites within one sequence.

4 Evolution of regulatory DNA

In statistical picture developed so far, background sequences and functional sites are reduced to ensembles P_0 and Q. This picture is incomplete in two ways. One one hand, it is quite disconnected from the biophysical aspects discussed before: the specific function of binding sites hardly enters the standard formalism of position weight matrices. On the other hand, there is not yet any notion of time and dynamics. Sequences change by various mutation processes, and the observed sequence ensembles derive from this evolutionary dynamics. The evolution of functional loci is fundamentally different from that of background sequence: it is subject to *natural selection*, that is, the fitness of an organism depends on its genotype $\mathbf{a}$ at a functional locus via the effects on the regulated gene. At this point, the biophysics of binding enters the evolution of functional sequences [24, 25, 26]. Moreover, it becomes clear that the statistical framework has to be extended from individual sequences to distributions of genotypes in a population. In this section, we develop an evolutionary picture of regulatory DNA, from which we obtain expressions for the sequence ensembles P_0, Q, and the score function S. The next four paragraphs are a self-contained introduction to the underlying concepts of population genetics.

Deterministic population dynamics and fitness. We start by describing the evolution of a large population, which contains individuals of different genotypes $\mathbf{a}$. Each genotype is assumed to produce a specific *phenotype*, which may influence the reproductive success of the individuals carrying it. With respect to factor binding, the phenotype can be associated with the binding energy $E(\mathbf{a})$, since presumably all organismic effects of a locus depend on its genotype only via the binding energy. However, the discussion in the following paragraphs is more general. For a more detailed presentation, see, e.g., ref. [27].

We first assume that the subpopulations of a given genotype reproduce separately, i.e., there neither transitions between genotypes through mutations nor (in a sexually reproducing population) mixing through genomic recombination. Writing the dynamics of the subpopulations in the form of simple growth laws,

$$\frac{\mathrm{d}}{\mathrm{d}t}N_{\mathbf{a}}(t) = F_{\mathbf{a}}(t)N_{\mathbf{a}}(t), \tag{27}$$

defines the (Malthusian) *fitness* $F_{\mathbf{a}}(t)$ of each genotype. For notational simplicity, we now limit ourselves to the case of just two genotypes **a** and **b**, where (27) can be written as growth laws for the total population size $N(t) \equiv N_{\mathbf{a}}(t) + N_{\mathbf{b}}(t)$ and for the population fraction $x(t) \equiv N_{\mathbf{b}}(t)/N(t)$ of genotype **b**,

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = \bar{F}(t)N(t), \tag{28}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \Delta F_{\mathbf{ab}}(t)\, x(t)[1 - x(t)], \tag{29}$$

with $\bar{F}(t) \equiv [1 - x(t)]F_{\mathbf{a}}(t) + x(t)F_{\mathbf{b}}(t)$ and $\Delta F_{\mathbf{ab}}(t) \equiv F_{\mathbf{b}}(t) - F_{\mathbf{a}}(t)$. This decomposition is useful since the overall growth rate $\bar{F}(t)$ is often strongly time-dependent due to external conditions (e.g., seasonality), while fitness differences, which reflect intrinsic properties of the phenotypes, are more stable. Different genotypes coexisting in a population frequently produce the same or very similar phenotypes and thus have equal fitness ($\Delta F_{\mathbf{ab}} = 0$).

Assuming $\Delta F_{\mathbf{ab}}$ to be constant over the time of observation, the solution of eq. (29) is the evolutionary trajectory

$$x(t) = \frac{x_0 \exp[\Delta F_{\mathbf{ab}}(t - t_0)]}{1 + x_0(\exp[\Delta F_{\mathbf{ab}}(t - t_0)] - 1)} \tag{30}$$

with the initial condition $x(t_0) = x_0$, shown in fig. 6(a). For $\Delta F_{\mathbf{ab}} \neq 0$, the fixed points of this dynamics are the monomorphic population states $x = 0$, and $x = 1$, of which $x = 1$ is stable for $\Delta F_{\mathbf{ab}} > 1$ and $x = 0$ for $\Delta F_{\mathbf{ab}} < 1$. The approach to the stationary state takes place on a characteristic time scale $\tau_d = 1/\Delta F_{\mathbf{ab}}$. In the important case of *neutral evolution* ($\Delta F_{\mathbf{ab}} = 0$), the evolutionary outcome remains indefinite. These results, which can readily be generalized to more than two phenotypes, are a simple version of Fisher's *fundamental theorem of natural selection*: any population with initially coexisting phenotypes of different fitness will evolve towards a state where only the fittest phenotype is present.

Fisher's theorem seems to prove the popularized Darwinian notion of the "survival of the fittest". However, it rests on very restrictive assumptions that are never fulfilled in a natural population. The deterministic growth law (29) neglects mutations and recombinations, as well as the reproductive fluctuations present in any population due to its finite number of individuals. These other evolutionary forces have to be incorporated in our theoretical picture before we can even define fitness as a measurable quantity and before the theory can address the important case of neutral evolution.

Stochastic dynamics and genetic drift. Stochastic fluctuations of the reproduction process in a large but finite population have been studied extensively in population genetics, see [28, 29]. They are called *genetic drift*, an

unfortunate name which may falsely suggest a deterministic effect. To take these fluctuations into account, we replace eq. (27) by a stochastic growth law,

$$\frac{\mathrm{d}}{\mathrm{d}t}N_{\mathbf{a}}(t) = F_{\mathbf{a}}(t)N_{\mathbf{a}}(t) + \chi_{\mathbf{a}}(t), \tag{31}$$

where $\chi_{\mathbf{a}}(t)$ are Gaussian random variables with $\overline{\chi_{\mathbf{a}}(t)} = 0$ and

$$\overline{\chi_{\mathbf{a}}(t)\chi_{\mathbf{b}}(t')} = N_{\mathbf{a}}(t)\,\delta(t-t')\,\delta_{\mathbf{a},\mathbf{b}}. \tag{32}$$

This form of noise is simply due to the law of large numbers, and the continuum dynamics (31) emerges as an effective large-N description for a plethora of discrete evolution models, which are defined at the level of individuals and have finite generation times. In the application to real populations, N has to be interpreted as the so-called *effective population size*, which can be inferred from genome data and is in general smaller than the actual population size.

In the case of two genotypes, eq. (31) can again be projected onto the population fraction x,

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \Delta F_{\mathbf{ab}}(t)\,x(t)[1-x(t)] + \chi_x(t), \tag{33}$$

where $\chi_x(t) = (\partial x/\partial N_{\mathbf{a}})\chi_{\mathbf{a}}(t) + (\partial x/\partial N_{\mathbf{b}})\chi_{\mathbf{b}}(t)$ are Gaussian random variables with zero mean and

$$\overline{\chi_x(t)\chi_x(t')} = \frac{x(1-x)}{N}\,\delta(t-t'). \tag{34}$$

This dynamics produces stochastic evolutionary trajectories $x(t)$ as shown in fig. 6(b). To capture their statistics, we convert the Langevin equation (33) into a Fokker-Planck equation for the probability distribution of the genotype composition [30, 28],

$$\frac{\partial}{\partial t}\mathcal{P}(x,t) = \frac{1}{2N}\frac{\partial^2}{\partial x^2}x(1-x)\mathcal{P}(x,t) - \Delta F_{\mathbf{ab}}(t)\frac{\partial}{\partial x}x(1-x)\mathcal{P}(x,t). \tag{35}$$

The mathematical subtlety of this equation lies in the x-dependent diffusion "constant" $x(1-x)/2N$, which reflects the multiplicative nature of the reproduction process. As a consequence, the two monomorphic population states $x = 0$ and $x = 1$ are also fixed points also of the stochastic dynamics. Any evolutionary trajectory $x(t)$ will eventually lead to one of these states with probability 1; this is called the *fixation* of the corresponding genotype in the population. In other words, the Fokker-Planck equation (35) describes diffusion in the interval $(0,1)$ with *absorbing boundaries*. There is a family of stationary states

$$\mathcal{P}(x) = (1-\phi)\delta(x) + \phi\delta(1-x), \tag{36}$$

parametrized by the *fixation probability* ϕ of genotype $\mathbf{b}$. The value of ϕ depends on the initial condition x_0 and can be computed by solving the backward diffusion equation

$$\frac{\partial}{\partial t}\mathcal{P}(x,t|x_0,t_0) = x_0(1-x_0)\left(\frac{1}{2N}\frac{\partial^2}{\partial x_0^2} - \Delta F_{\mathbf{ab}}(t)\frac{\partial}{\partial x_0}\right)\mathcal{P}(x,t|x_0,t_0). \quad (37)$$

For time-independent $\Delta F_{\mathbf{ab}}$, the stationary solution $\phi(x_0) \equiv \lim_{t\to\infty}\mathcal{P}(x = 1,t|x_0,t_0)$ has the form [30, 28]

$$\phi(x_0,\Delta F_{\mathbf{ab}},N) = \frac{1-\exp(-2N\Delta F_{\mathbf{ab}}x_0)}{1-\exp(-2N\Delta F_{\mathbf{ab}})}, \quad (38)$$

which for near-neutral evolution ($N\Delta F_{\mathbf{ab}} \ll 1$) reduces to

$$\phi(x_0,0,N) = x_0 + N\Delta F_{\mathbf{ab}}\, x_0(1-x_0) + \ldots. \quad (39)$$

The characteristic time τ_s of the stochastic dynamics interpolates between the diffusive scale N and the deterministic scale: $\tau_s \approx \min(N,\tau_d)$. It determines the typical time of the evolution process up to fixation, shown shaded in fig. 6(b).

Hence, the stochastic population dynamics depends no longer only on the fitness difference of the genotypes as in the deterministic case, but also on the initial state of the population and the the population size. Yet, our evolutionary picture is still incomplete. Population states with coexisting genotypes enter the dynamics as initial conditions, but since mutations are neglected, the model does not explain how this coexistence is generated and maintained.

Mutation processes and evolutionary equilibria. At the level of an individual, mutations are rare stochastic genotype changes $\mathbf{a} \to \mathbf{b}$, which take place with rates $\mu_{\mathbf{a}\to\mathbf{b}}$, often coupled to the reproduction process. (These rates are all of the same order of magnitude, in estimates we therefore omit the indices.) We include mutations into the population dynamics (31) by their systematic effect on the genotype subpopulations,

$$\frac{\mathrm{d}}{\mathrm{d}t}N_{\mathbf{a}}(t) = F_{\mathbf{a}}(t)N_{\mathbf{a}}(t) + \sum_{\mathbf{b}}\left[\mu_{\mathbf{b}\to\mathbf{a}}N_{\mathbf{b}}(t) - \mu_{\mathbf{a}\to\mathbf{b}}N_{\mathbf{a}}(t)\right] + \chi_{\mathbf{a}}(t), \quad (40)$$

while their stochastic effect (whose variance is of order $N\mu$) is neglected since it is small against the reproductive sampling noise $\chi_{\mathbf{a}}(t)$. In the case of two different genotypes, this dynamics can again be projected onto the variable x,

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = \Delta F_{\mathbf{ab}}(t)\, x(t)[1-x(t)] + \mu_{\mathbf{a}\to\mathbf{b}}[1-x(t)] - \mu_{\mathbf{b}\to\mathbf{a}}\, x(t) + \chi_x(t), \quad (41)$$

which leads to the Fokker-Planck equation [31]

$$\begin{aligned}\frac{\partial}{\partial t}\mathcal{P}(x,t) = {} & \frac{1}{N}\frac{\partial^2}{\partial x^2}x(1-x)\mathcal{P}(x,t) - \Delta F_{\mathbf{ab}}(t)\frac{\partial}{\partial x}x(1-x)\mathcal{P}(x,t) \\ & -\mu_{\mathbf{a}\to\mathbf{b}}\frac{\partial}{\partial x}(1-x)\mathcal{P}(x,t) + \mu_{\mathbf{b}\to\mathbf{a}}\frac{\partial}{\partial x}x\mathcal{P}(x,t). \end{aligned} \quad (42)$$

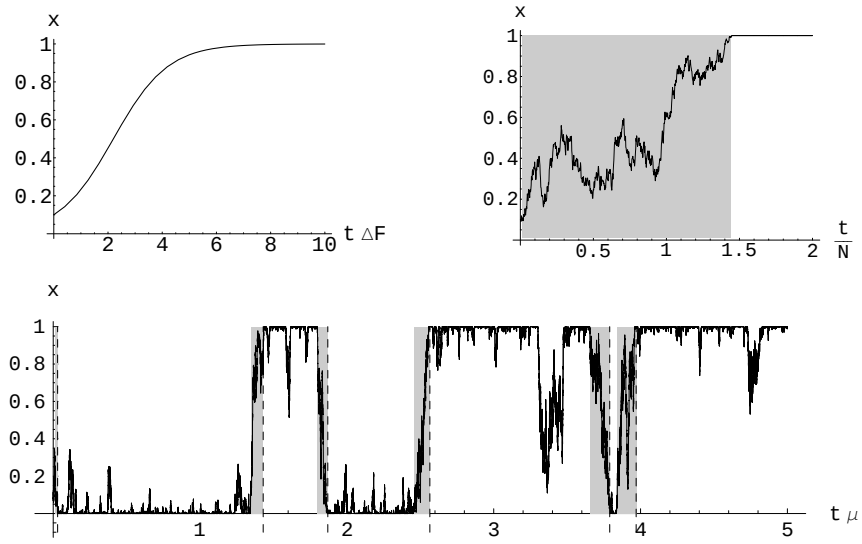

Fig. 6. Evolution of genotype composition $x(t)$**.** (a) Deterministic evolution with fitness difference $\Delta F_{\mathbf{ab}} > 0$, leading to certain fixation of genotype **b** (time is shown in units of $\tau_d = 1/\Delta F_{\mathbf{ab}}$). (b) Stochastic evolution with selection and genetic drift, leading to fixation of one of the genotypes. The time to fixation (grey shading) is of order τ_s ($N\Delta F_{\mathbf{ab}} = 0.5$, time is shown in units of N). (c) Stochastic evolution with selection, genetic drift, and mutations in the regime $N\mu \ll 1$, leading to a substitution dynamics with rates $u_{\mathbf{a}\to\mathbf{b}}$ and $u_{\mathbf{b}\to\mathbf{a}}$ given by (46). Substitution events are marked by dashed lines. The typical time between initial mutation and fixation (grey shading) for a given substitution, τ_s, is much shorter than the time between subsequent substitutions, $1/u_{\mathbf{a}\to\mathbf{b}}$ resp. $1/u_{\mathbf{b}\to\mathbf{a}}$ ($N\Delta F_{\mathbf{ab}} = 0.5$, $N\mu = 0.05$, time is shown in units of $1/\mu$).

For time-independent $\Delta F_{\mathbf{ab}}$, this equation has a single stable stationary state,

$$\mathcal{P}(x) = \frac{1}{Z}\, x^{-1+N\mu_{\mathbf{a}\to\mathbf{b}}} (1-x)^{-1+N\mu_{\mathbf{b}\to\mathbf{a}}} \exp(2N\Delta F_{\mathbf{ab}}\, x) \tag{43}$$

with a normalization constant Z that can be expressed in terms of Bessel and Gamma functions [32].

Substitution dynamics. Here we are interested in the stochastic evolution (42) and its equilibrium state (43) for $N\mu \ll 1$, which is the relevant dynamical regime for nuclear DNA in eukaryotes and in most prokaryotes (but not in viral systems). In this regime, the mutation term in (42) is small against the diffusion term except for values of x close to the boundaries 0 or 1. In this region, the continuum approximation of eq. (42) is no longer valid, and (43) has to be replaced by a stationary solution $\mathcal{P}_d(N_{\mathbf{a}})$ of the underlying discrete evolution model, which gives the probability that the population contains $N_{\mathbf{a}}$ individuals of genotype **a** (with $N_{\mathbf{a}} = N - N_{\mathbf{b}} = 0, 1, \ldots, N$). The discrete

solution is easily shown to have the singularity $\mathcal{P}_d(0) \simeq (N\mu_{\mathbf{a}\to\mathbf{b}})^{-1} P_d(1)$. This singularity is correctly captured if we use the approximation $P_d(N_\mathbf{a}) \simeq \int_{N_\mathbf{a}/N}^{(N_\mathbf{a}+1)/N} dx\, \mathcal{P}(x)$ for all $N_\mathbf{a}$ (except at the other boundary, where there is a similar singularity $\mathcal{P}_d(N) \simeq (N\mu_{\mathbf{b}\to\mathbf{a}})^{-1} P_d(N-1)$) [33].

From this solution, we read off the following characteristics of the evolutionary dynamics at equilibrium, which are illustrated by the trajectory of fig. 6(c) [32]:

(a) For sufficiently small values of μ, the population remains monomorphic for most of the time. Using the shorthands $Q(\mathbf{a}) \equiv \mathcal{P}_d(N_\mathbf{a} = 0)$ and $Q(\mathbf{b}) \equiv \mathcal{P}_d(N_\mathbf{a} = N)$, we have

$$Q(\mathbf{a}) + Q(\mathbf{b}) = 1 - O(\mu N \log N). \tag{44}$$

(b) The ratio of probabilities for the two monomorphic population states is given by the ratio of "forward" and "backward" mutation rate, the fitness difference, and the effective population size:

$$\frac{Q(\mathbf{b})}{Q(\mathbf{a})} = \frac{\mu_{\mathbf{a}\to\mathbf{b}}}{\mu_{\mathbf{b}\to\mathbf{a}}} \exp(2N\Delta F_{\mathbf{ab}}) + O(N\mu). \tag{45}$$

(c) The monomorphic population states $x = 0$ and $x = 1$ are unstable due to mutations even at arbitrarily small values of μ, which cause occasional transitions of the entire population from genotype $\mathbf{a}$ to $\mathbf{b}$, and vice versa. These so-called *substitutions* are marked by dashed lines in fig. 6(c). The substitution rate $u_{\mathbf{a}\to\mathbf{b}}$ can be evaluated as the product of creating a single mutant of genotype $\mathbf{b}$ in an initially monomorphic $\mathbf{a}$ population, $N\mu_{\mathbf{a}\to\mathbf{b}}$, and its probability of fixation, $\phi(x_0 = 1/N, \Delta F_{\mathbf{ab}}, N)$. The time between initial mutation and fixation (shown by grey shading in fig. 6(c)) is still of order τ_s and thus much shorter than the time scale $1/\mu$, on which mutation effects become important. Hence, the fixation probability ϕ is given to leading order by (38), which has been derived for $\mu = 0$. Together we have [30, 28]

$$u_{\mathbf{a}\to\mathbf{b}} = N\mu_{\mathbf{a}\to\mathbf{b}} \frac{1 - \exp(-2\Delta F_{\mathbf{ab}})}{1 - \exp(-2N\Delta F_{\mathbf{ab}})}. \tag{46}$$

Hence, the substitution rate $u_{\mathbf{a}\to\mathbf{b}}$ is enhanced over $\mu_{\mathbf{a}\to\mathbf{b}}$ for $\Delta F_{\mathbf{ab}} > 0$ and suppressed for $\Delta F_{\mathbf{ab}} < 0$, as shown in Fig. 7. For weak selection ($N|\Delta F_{\mathbf{ab}}| \ll 1$), eq. (46) becomes

$$u_{\mathbf{a}\to\mathbf{b}} = \mu_{\mathbf{a}\to\mathbf{b}}(1 + N\Delta F_{\mathbf{ab}} + \ldots). \tag{47}$$

This reproduces Kimura's famous original result: for neutral evolution, the substitution rate equals the mutation rate in an individual, independently of the population size. For this reason, the rates $\mu_{\mathbf{a}\to\mathbf{b}}$ are referred to as neutral mutation rates. For strong selection ($N|\Delta F_{\mathbf{ab}}| \gg 1 \gg |\Delta F_{\mathbf{ab}}|$), eq. (46) takes the asymptotic forms

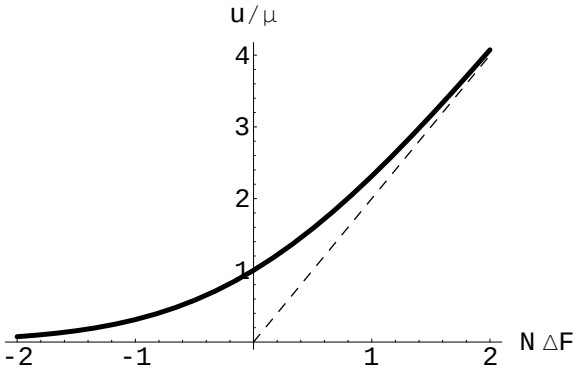

Fig. 7. Substitution rate in a population versus mutation rate in an individual. The ratio of these rates, $u_{\mathbf{a}\to\mathbf{b}}/\mu_{\mathbf{a}\to\mathbf{b}}$, depends on the product $N\Delta F_{\mathbf{ab}}$ of effective population size and fitness difference between the genotypes (in the relevant regime $N \gg 1$, $\Delta F_{\mathbf{ab}} \ll 1$, $N\Delta F_{\mathbf{ab}}$ finite). The substitution rate $u_{\mathbf{a}\to\mathbf{b}}$ is equal to $\mu_{\mathbf{ab}}$ for neutral mutations ($\Delta F_{\mathbf{ab}} = 0$), reduced for deleterious mutations ($\Delta F_{\mathbf{ab}} < 0$), and enhanced for advantageous mutations ($\Delta F_{\mathbf{ab}} > 0$).

$$u_{\mathbf{a}\to\mathbf{b}} = \mu_{\mathbf{a}\to\mathbf{b}} \begin{cases} 2N|\Delta F_{\mathbf{ab}}| \exp(2N\Delta F_{\mathbf{ab}}) & (2N\Delta F_{\mathbf{ab}} \ll 1), \\ 2N\Delta F_{\mathbf{ab}} & (2N\Delta F_{\mathbf{ab}} \gg 1). \end{cases} \tag{48}$$

The backward substitution rate $u_{\mathbf{b}\to\mathbf{a}}$ is given by a formula similar to (46) with $\Delta F_{\mathbf{ba}} = -\Delta F_{\mathbf{ab}}$. Forward and backward substitution rate have the simple ratio

$$\frac{u_{\mathbf{a}\to\mathbf{b}}}{u_{\mathbf{b}\to\mathbf{a}}} = \frac{\mu_{\mathbf{a}\to\mathbf{b}}}{\mu_{\mathbf{b}\to\mathbf{a}}} \exp(2N\Delta F_{\mathbf{ab}}) \tag{49}$$

for $N \gg 1$. Comparing with (45), we obtain the consistency condition

$$\frac{u_{\mathbf{a}\to\mathbf{b}}}{u_{\mathbf{b}\to\mathbf{a}}} = \frac{Q(\mathbf{b})}{Q(\mathbf{a})}. \tag{50}$$

Hence, for sufficiently small mutation rates ($\mu N \log N \ll 1$), a simple picture emerges: The evolution of a population can be described as a sequence of transitions between monomorphic genotype states (substitutions). The substitution rate u is determined by the corresponding mutation rate in an individual, the fitness difference between the genotypes, and the effective population size.

Neutral dynamics in sequence space, sequence entropy. This evolutionary picture can be generalized to multiple genotypes, for example, the 4^ℓ dimensional sequence space of genomic loci $\mathbf{a} = (a_1, \ldots, a_\ell)$. Transitions between different sequence states are point mutations $\mathbf{a} \to \mathbf{b}$, which change exactly one nucleotide. (We neglect here insertion and deletion processes, which change the length of the sequence). We first discuss neutral evolution,

where the substitution rate $u_{\mathbf{a}\to\mathbf{b}}$ equals the mutation rate in an individual, $\mu_{\mathbf{a}\to\mathbf{b}}$, for all elementary transitions $\mathbf{a} \to \mathbf{b}$. Bona fide neutral mutation rates can be inferred from DNA sequence alignments of sufficiently close species, recent insights have also come from studying repeat elements.

We assume the neutral dynamics has an equilibrium distribution $P_0(\mathbf{a})$ which obeys *detailed balance*, i.e., the relation

$$\frac{\mu_{\mathbf{a}\to\mathbf{b}}}{\mu_{\mathbf{b}\to\mathbf{a}}} = \frac{P_0(\mathbf{b})}{P_0(\mathbf{a})} \tag{51}$$

holds for each pair of sequence states linked by an elementary transition process $\mathbf{a} \to \mathbf{b}$. This says that the probability current at equilibrium, $\mu_{\mathbf{a}\to\mathbf{b}}P_0(\mathbf{a}) - \mu_{\mathbf{b}\to\mathbf{a}}P_0(\mathbf{b})$, vanishes for each elementary transition. Clearly, any distribution $P_0(\mathbf{a})$ satisfying the conditions (51) is stationary under the dynamics with rates $\mu_{\mathbf{a}\to\mathbf{b}}$, but not every such dynamics has a stationary distribution which satisfies (51) (the simplest counterexample involving three states and a circular probability current $\mathbf{a} \to \mathbf{b} \to \mathbf{c}$ at stationarity). However, as will be verified below, detailed balance is a good approximation for the genomic substitution dynamics at least in prokaryotes. (There are known violations at CpG islands in eukaryotes [34]). In the simplest type of models, every nucleotide a mutates independently of all other positions with uniform rates $\mu_{a\to b}$ (i.e., $\mu_{\mathbf{a}\to\mathbf{b}} = \mu_{a\to b}$ for any two sequences $\mathbf{a} = (\ldots, a, \ldots)$ and $\mathbf{b} = (\ldots, b, \ldots)$ differing by exactly one nucleotide). This produces a factorized equilibrium distribution $P_0(\mathbf{a})$ of the form (12).

We can project the equilibrium distribution onto a measurable quantity as independent variable. For binding site sequences, a convenient choice is the binding energy E, and the projected distribution $P_0(E)$ has the form (13). Hence we can define the *sequence entropy* [35]

$$S_0(E) = \log P_0(E), \tag{52}$$

which counts the log density of sequence states $\mathbf{a}$ at energy E, weighed by the distribution $P_0(\mathbf{a})$.

Dynamics under selection, the score-fitness relation. The dynamics of substitutions can be studied in the same way for evolution under selection, which is specified at the level of genotypes by an arbitrary fitness function $F(\mathbf{a})$ [37, 17]. This generalizes the results of [36] for a model with selection acting independently at different nucleotide positions, i.e., $F(\mathbf{a}) = \sum_{i=1}^{\ell} f_i(a_i)$. For each elementary transition $\mathbf{a} \to \mathbf{b}$, the substitution rate $u_{\mathbf{a}\to\mathbf{b}}$ is determined by the neutral rate $\mu_{\mathbf{a}\to\mathbf{b}}$, the fitness difference $\Delta F_{\mathbf{ab}}$, and the effective population size N according to (46). Given the detailed balance (51) of neutral evolution and the relation (49) between forward and backward rates, it then follows immediately that the evolutionary dynamics under selection also obeys detailed balance, as given by (50) with an equilibrium distribution $Q(\mathbf{a})$ of the form (45). Thus we have [37, 17]:

The equilibrium distribution $Q(\mathbf{a})$ of fixed genotypes generated by a substitution dynamics (46) *with fitness function $F(\mathbf{a})$ is related to its neutral counterpart $P_0(\mathbf{a})$ by*

$$Q(\mathbf{a}) = P_0(\mathbf{a}) \exp[2NF(\mathbf{a}) + \text{const.}], \tag{53}$$

with the constant given by normalization.

We can project eq. (53) onto the fitness as independent variable. Defining the distribution $Q(F) \equiv \sum_{\mathbf{a}} Q(\mathbf{a})\delta(F(\mathbf{a}) - F)$, similarly $P_0(F)$, and the sequence entropy $S_0(F) \equiv \log P_0(F)$, the projected identity takes the form

$$Q(F) = \exp[2NF + S_0(F) + \text{const.}] \tag{54}$$

For binding site sequences, we have a similar projection on the binding energy, $Q(E) = \exp[2NF(E) + S_0(E) + \text{const.}]$, since all genotypes with the same "phenotype" E have the same fitness, i.e., the same score S. The projected identities express the equilibrium distribution under selection in terms of fitness and sequence entropy, reflecting the balance between stochasticity (genetic drift) and selection [17]. For strong selection, the exponent $2NF - S_0$ is dominated by the fitness term, and $Q(F)$ takes appreciable values only at points of near-maximal fitness, i.e., where $F_{\max} - F \lesssim 1/2N$. For moderate selction, there is a nontrivial balance between both terms, and for weak selection, the Q distribution can be approximated by its neutral counterpart $P_0 = \exp(S_0)$. Clearly, the roles of fitness and sequence entropy are formally analogous to those of energy and entropy in statistical physics of thermodynamic systems, if $2N$ is identified with the inverse temperature $1/k_BT$. Some consequences of this analogy are discussed in ref. [38].

The dynamics of substitutions establishes a rather general evolutionary grounding of genome statistics, if we identify the equilibrium distributions $P_0(\mathbf{a})$ and $Q(\mathbf{a})$ with the genomic distributions discussed in the previous section, as already anticipated by our notation. Comparing eqs. (53) and (14) gives a relation between fitness and score [17, 15]:

The log-likelihood score $S(\mathbf{a}) = \log[Q(\mathbf{a})/P_0(\mathbf{a})]$ equals the fitness function multiplied by twice the effective population size up to a constant,

$$S(\mathbf{a}) = 2NF(\mathbf{a}) + \text{const.}. \tag{55}$$

This relation allows us to use sequence data of a given genome to infer quantitative patterns of its evolution. We now discuss specific consequences for the evolution of regulatory DNA; an application to protein evolution can be found in ref. [36].

Measuring selection for binding sites. We first give a precise definition of functionality for regulatory (and other) elements: A binding *locus* is functional if the genotype at that locus is under selection (for binding of the corresponding factor). Nonfunctional loci have evolutionarily neutral genotypes. This definition asks whether binding at a given locus makes a difference

to the organism or not. It is weaker than that of a functional *binding site*, which is a functional locus with a sequence **a** that is likely to actually bind the factor. A functional locus can lose its binding sequence due to deleterious mutations, leading to suboptimal fitness of the organism. Conversely, a non-functional locus can have by chance a sequence which does bind the factor: this is a spurious binding site without consequences for the organism.

To measure the selection on functional sites *in silico*, we apply the identity (55) to the genomic distributions $P_0(\mathbf{a})$ and $Q(\mathbf{a})$. (Assuming equilibrium for most loci seems to be justified for our example of CRP binding sites in *E. coli* since we find very similar distributions in the distant bacterial species *Salmonella typhimurium*, and the factor protein itself is highly conserved between these species.) After projection onto the energy, the fitness landscape $2NF(E)$ for CRP binding sites is thus given by fig. 4(b) [15]. The fitness is constant in the no-binding region ($E \gtrsim E_s \approx 13$) since the evolution is always neutral in that region. This constant is set to 0 in our normalization, i.e., $F(E)$ measures the fitness gain of functional sites due to factor binding. Loci with strong binding are also under strong selection, with effective fitness values $2NF$ of order 10. Genetic drift counteracts selection, producing also loci with weaker binding and reduced effective fitness. This fitness "landscape" is thus qualitatively of the form predicted from the underlying biophysics [24, 17]. Of course, it should be kept in mind that this landscape results from averaging over a family of binding sites, which may have a spectrum of individual selection coefficients and selected binding strengths.

Nucleotide frequency correlations. A further consequence of (54) is the generic occurence of nucleotide frequency correlations within functional loci [17]. If the fitness function $F(\mathbf{a})$ is not additive in the nucleotide positions, nucleotide frequencies are correlated in selected genotypes even if they are independent under neutral evolution. This happens quite generically since selection acts on the entire genotype **a** as a functional unit and not on its single nucleotides. For binding sites, fitness effects follow from the expression level of the regulated gene, which depends on the sequence **a** via the binding probability of the corresponding transcription factor. While the binding energy is often approximately additive in the nucleotide positions as given by (1), the binding probability (10) is a strongly nonlinear function of the energy. This introduces correlations between nucleotide frequencies at *any two* positions within functional loci, preventing factorization of the distribution $Q(\mathbf{a})$.

Stationary evolution of binding sites. Functional loci with a substantial level of selection (as found for the CRP binding sites in *E. coli*) evolve in a way quite different from background sequence. This is quantified in fig. 8(a), which shows pairs of binding energies (E_1, E_2) for experimentally verified CRP binding sites in *E. coli* and the corresponding sites regulating orthologous genes in *S. typhimurium* [26, 15]. The evolutionary distance t between the two species and characteristics of the neutral mutation process can be in-

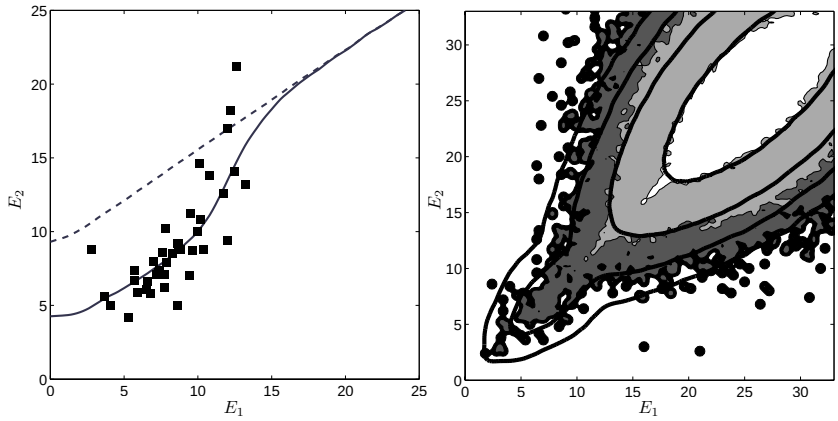

Fig. 8. Evolution of binding sites. (a) Binding energy pairs (E_1, E_2) for 32 experimentally verified CRP binding sites in *E. coli* from the DPInteract database [41] and their aligned orthologs in *S. typhimurium* (dots). Conditional expectation value for the binding energy in *S. typhimurium* under neutral evolution, $\langle G_0(E_2|E_1)\rangle$ (dashed line), and under selection, $\langle G_f(E_2|E_1)\rangle$ (solid line). (b) Distribution of energy pair counts $W_{\rm dat}(E_1, E_2)$ (filled contours), compared to the distribution $W(E_1, E_2)$ given by the Bayesian model (59). The symmetry of these distributions under exchange of E_1 and E_2 reflects detailed balance of the substitution dynamics. From [15, 39].

ferred from alignments of background sequence. The "phenotypic" evolution of CRP binding is quantified by the *energy transition probabilities* $G_0(E_2|E_1)$ under neutral evolution and $G_f(E_2|E_1)$ under stationary selection [15]. These are readily obtained by simulating the substitution dynamics over a time interval t for given initial value E_1, both with neutral rates $\mu_{\mathbf{a}\to\mathbf{b}}$ and with rates $u_{\mathbf{a}\to\mathbf{b}}$ given by (46) and the fitness function $2NF(E)$ measured in *E. coli*. The resulting conditional expectation values $\langle G_0(E_2|E_1)\rangle$ and $\langle G_f(E_2|E_1)\rangle$ for the binding energy in *S. typhimurium* are also shown in fig. 8(a). The data conform to the selection model, showing a substantially stronger conservation of binding energy than expected for neutral evolution [26, 15, 39].

We can now build a probabilistic model for cross-species comparisons [15]. It is based on the joint distributions of energy pairs

$$P_0(E_1, E_2) = G_0(E_2|E_1)\, P_0(E_1) \tag{56}$$

under neutral evolution and

$$Q(E_1, E_2) = G_f(E_2|E_1)\, Q(E_1) \tag{57}$$

under stationary selection, which are determined by the corresponding distributions in one species and the energy transition probabilities. Detailed balance of the substitution dynamics implies

$$\frac{P_0(E_2)}{P_0(E_1)} = \frac{G_0(E_2|E_1)}{G_0(E_1|E_2)} \quad \text{and} \quad \frac{Q(E_2)}{Q(E_1)} = \frac{G_f(E_2|E_1)}{G_f(E_1|E_2)}, \tag{58}$$

i.e., the joint distributions $P_0(E_1, E_2)$ and $Q(E_1, E_2)$ must be symmetric functions of their arguments. These distributions combine into a model for pairs of aligned loci, which generalizes the single-species model (22) and takes the form

$$W(E_1, E_2) = (1 - \lambda)P_0(E_1, E_2) + \lambda Q(E_1, E_2). \tag{59}$$

(This model can be extended further to include non-stationary selection.) The distribution $W(E_1, E_2)$ with a fraction of functionality $\lambda = 0.0018$ is in excellent agreement with the count distribution $W_{\mathrm{dat}}(E_1, E_2)$ obtained from *E. coli* and *S. typhimurium*, as shown in fig. 8(b). The symmetry of W_{dat} thus corroborates the underlying assumption of detailed balance. Analogous Bayesian models can be defined for more than two species related by a phylogeny. This approach has been applied to binding site prediction in bacteria [15]; a related study of several species of funghi has been reported in ref. [40].

Adaptive evolution of binding sites. What does this picture say about the adaptive evolution of transcriptional regulation in response to a newly arising selection pressure? The evolution from a genotype with marginal binding ($E(\mathbf{a}) \approx E_s$) to strong binding requires only about three uphill point mutations in the fitness landscape of fig. 4(b), i.e., there is an effective fitness gain $2N\Delta F \approx 3$ per mutation. Hence, according to (48), the rate of uphill substitutions per locus is enhanced by a factor $2N\Delta F \cdot d(\mathbf{a}, \mathbf{a}^*)$ at least of order 10 over the neutral point mutation rate per nucleotide. At the same time, the downhill rate is strongly suppressed. This shows that the adaptive formation of a binding site from background sequence can indeed be a rapid mode of regulatory evolution, due to the substantial level of selection [17].

However, this mode is only efficient if adaptation can set in immediately after the selection pressure is established. In larger regulatory regions, the exact position of a binding site is often not important. We assume the initial genome contains a set of $\tilde{L}$ *shadow sites*, i.e., positions $r_1, \ldots, r_{\tilde{L}}$ where a given sequence $\mathbf{a}$ would have the same regulatory effect. If one of these shadow sites has already a genotype with marginal binding, it acts as a "seed" for the onset of adaptation [42]. On the other hand, if all shadow sites of the initial genome have energy $E > E_s$, there is typically a substantial waiting time of neutral evolution before one of them reaches the threshold energy E_s. Assuming the initial genome to be entirely background sequence, it will contain at least one such seed if $\int_{E<E_s} P_0(E)\mathrm{d}E \gtrsim 1/\tilde{L}$, which is a joint condition on $\tilde{L}$ and the site length ℓ: the shadow regulatory region must be long enough and binding sites must be short enough. The example shows that the evolvability of regulation imposes constraints on genome architecture [17].

5 Towards a dynamical picture of the genome

The relationship $S = 2NF + \text{const.}$ between score and fitness is a cornerstone of the theoretical picture developed so far, which links its population genetic, bioinformatic and biophysical arches. It relates a key evolutionary variable with the statistics of genomic frequency counts. The physical binding energy is an appropriate phenotypic variable on which fitness and score depend, because molecular function is determined by binding interactions.

We have discussed this picture for transcription factor binding sites, but it can be applied more generally to functional elements in genomes. It relates the statistics of these elements in one genome with their evolutionary dynamics, which is observed in cross-species comparisons. This dynamics is shaped by selection: The components of functional elements are coupled by a common fitness function. Hence, *functional correlations lead to evolutionary correlations*. These can be traced in the Q distribution over fixed genomes of a functional element; other methods use the statistics of polymorphisms within a population.

Thus, the picture of the genome as a system with multiple interactions has a fundamental dynamical significance. This is important since it allows us to trace functional modules from evolutionary patterns. We conclude the article with a brief outlook on various levels of functional integration for regulatory sequences.

Evolutionary interactions between sites. Regulatory function is often determined not by single binding sites, but jointly by a group of sites in the same regulatory region [43]. An important mechanism is *binding cooperativity*, i.e., the formation of a protein complex between two (or more) factors bound to their corresponding DNA sites. The binding energy of this complex has the form $E = E_1 + E_2 + \Delta E_{12}$, where E_1 and E_2 are the energies of the factors bound individually and $\Delta E_{12} < 0$ is the energy gain due to the protein-protein interaction, which is of the order of a few $k_B T$. Cooperative binding has a number of functional effects [1]:

(a) It increases the signal-to-noise ratio for the targeting of regulatory input to a specific gene, which is important in larger eukaryotic genomes, where single spurious binding sites are abundant in background sequence.

(b) It sharpens the response of the binding probability to variations in the factor concentrations around their threshold value. This follows from the thermodynamics of two factors, which is a straightforward generalization of the case of a single factor discussed in section 2.

(c) It implements logical connections between regulatory input signals to a given gene. The simplest example is an AND connection between two factors, where the regulated gene is affected only if both factors are simultaneously present. This happens if the binding energies and factor concentrations are such that individual binding is weak but joint binding is strong. Larger groups

of binding sites can encode a whole repertoire of more complicated logical functions [44].

Regulatory modules with several jointly acting binding sites are frequently found in eukaryotes. The functional coupling of sites in a module translates into interactions between these sites in their sequence evolution. The genomic functional element, i.e., the subset of the regulatory region on which selection acts, is the module as a whole. Its fitness $F(E_1, E_2, \Delta E_{12}, \ldots)$ is a joint function of the binding energies as the relevant phenotypic variables [24, 17]. The evolutionary dynamics under this selection allows for a large number of *compensatory changes*, i.e., pairs of correlated substitutions changing two binding energies such that the fitness remains constant. These lead to nucleotide frequency correlations between different sites. Such compensatory changes have indeed been observed in experiments on *Drosophila* promoters [45].

Site-shadow interactions. In larger regulatory regions, there is a number of shadow sites where a binding sequence **a** would have a similar regulatory effect as at the functional sites present. In that case, the genomic functional element contains not only the functional binding sites but also the shadow sites. Once a functional site has disappeared due to deleterious mutations, a shadow site can turn functional by adaptive evolution as described in the last section. The resulting evolutionary dynamics leads to sequence turnover with the actual binding sites present at different but functionally equivalent positions [37]. Substantial sequence turnover has been observed in a number of case studies [46, 45, 47, 48, 49, 50]. Also the number of actual sites is subject to evolutionary variation since the same regulatory effect, i.e., the same fitness, can be distributed over fewer stronger or more weaker sites. With increasing number $\tilde{L}$ of shadow positions, one expects that the number of actual sites grows while individual sites get weaker [37].

Gene interactions. Evolutionary interactions are not limited to regulatory elements for the same gene. An example are gene duplications and the subsequent evolution of the daughter genes. Selection acts jointly on this pair of genes [51], which have initially identical functions, eventually leading to either loss of one of them or to *subfunctionalization*, which has been argued to be an important mode of genome evolution in eukaryotes [52, 53]. This process can take place by regulation, i.e., via a correlated distribution of the regulatory elements on the daughter genes. More generally, the evolution of genes in a regulatory network is correlated if their functions are coupled either in series (i.e., one gene acts on the other) or in parallel (i.e., they are part of alternative pathways for the same function). Although some regulatory networks in model organisms – e.g. the embryonic development in the sea urchin [54] – have been studied in detail, we lack a coherent view of their functional evolution to date.

Evolutionary innovations. Under stationary selection, functional elements are more conserved than background sequence, and the score-fitness relation

quantifies the amount of conservation. But evolution is, of course, not limited to conservation. On one hand, there is typically a multitude of different genotypes yielding the same molecular function, and the evolutionary dynamics continuously plays with these alternatives. On the other hand, organisms face long-term changes of their environment, which lead to new selection pressures and a response by adaptive evolution of *new functions*. If regulation is to account for a large part of the diversification in higher eukaryotes, loss or gain of regulatory function should be an important mode of molecular evolution. Changes in regulatory DNA leading to new functions of gene networks have been observed [55], and it is possible to extend the statistical models described in the previous section to include evolutionary gain or loss of function of individual binding sites [15]. On a broader scale, understanding the molecular basis of evolutionary innovations is a major challenge for theory and experiment in the coming years. It will profoundly change our dynamical view of the genome.

References

1. M. Ptashne and A. Gann, *Genes and Signals*, Cold Spring Harbour Laboratory Press (2002).
2. D. Tautz, Current Opinion in Genetics & Development **10**, 575-79 (2000).
3. G.A. Wray, M.W. Hahn, H. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and R.A. Romano, Mol. Biol. Evol. **20**, 1377-1419 (2003).
4. O.G. Berg, R.B. Winter, and P.H. von Hippel, Biochemistry **20**, 6929-48 (1981).
5. R.B. Winter and P.H. von Hippel, Biochemistry **20**, 6948-60 (1981).
6. R.B. Winter, O.G. Berg, and P.H. von Hippel, Biochemistry **20**, 6961-77 (1981).
7. P.H. von Hippel and O.G. Berg, Proc. Natl. Acad. Sci. **83**, 1608-1612 (1986).
8. A. Sarai and Y. Takeda, Proc. Natl. Acad. Sci. **86**, 6513-17 (1989).
9. D. Fields, Y. He, A. Al-Uzri, and G. Stormo, J. Mol. Biol. **271**, 178-94 (1997).
10. G.D. Stormo and D.S. Fields, Trends Biochem. Sci. **23**, 109-13 (1998).
11. M. Oda, K. Furukawa, K. Ogata, A. Sarai, and N. Nakamura, J. Mol. Biol. **276**, 571-90 (1998).
12. K. Omagari et al., FEBS Letters **563**, 55-58 (2004).
13. U. Gerland, D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. **99**, 12015-20 (2002).
14. M. Djordjevic, A.M. Sengupta, and B.I. Shraiman, Genome Res. **13**, 2381-90 (2003).
15. V. Mustonen and M. Lässig, Proc. Natl. Acad. Sci. **102**, 15936-41 (2005).
16. M. Slutsky and L.A. Mirny, Bioinformatics **15**, 2539-40 (2002).
17. J. Berg, S. Willmann, and M. Lässig, BMC Evolutionary Biology **4**, 42 (2004).
18. R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, Cambridge University Press (1998).
19. G. Stormo and G.W. Hartzell, Proc. Natl. Acad. Sci. **86**, 1183-7.
20. G. Hertz and G. Stormo, Bioinformatics **15**, 563-77 (1999).
21. N. Rajewsky, N.D. Socci, M. Zapotocky, and E.D. Siggia, Genome Res. **12**, 298-308 (2002).

22. E. van Nimwegen, M. Zavolan, N. Rajewsky, and E.D. Siggia, Proc. Natl. Acad. Sci. **99**, 7323-28 (2002).
23. B. Lenhard, A. Sandelin, L. Mendoza, P. Engström, N. Jareborg, and W.W. Wasserman, Jour. Biol. **2**, 13 (2003).
24. U. Gerland and T. Hwa, J. Mol. Evol. **55**, 386-400 (2002).
25. A.M. Moses, D.Y. Chiang, M. Kellis, E.S. Lander, and M.B. Eisen, BMC Evol Biol. **3**, 19 (2003).
26. C.T. Brown and C.G. Callan, Proc. Natl. Acad. Sci. **101**, 2404-09 (2003).
27. J. Hofbauer and K. Sigmund, *The Theory of Evolution and Dynamical Systems*, Cambridge University Press (1988).
28. M. Kimura and J.F. Crow, *An Introduction to Population Genetics Theory*, Harper & Row (New York, 1973).
29. M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press (1983).
30. M. Kimura, Genetics **47**, 713-19 (1962).
31. M. Kimura and T. Ohta, Genetics **61**, 763-71 (1969).
32. M. Rouzine, A. Rodrigo, and J.M. Coffin, Microbiol. Mol. Biol. Reviews **65**, 151-85 (2001).
33. D. Grün and M. Lässig, to be published.
34. P. Arndt and T. Hwa, Bioinformatics **21**, 2322-28 (2005).
35. L. Peliti, Europhys. Lett. **57**, 745-51 (2002).
36. A.L. Halpern and W.J. Bruno, Mol. Biol. Evol. **15**, 910-17 (1998).
37. J. Berg and M. Lässig, Biophysics (Moscow) **48**, Suppl. 1, 36-44 (2003).
38. G. Sella and A. Hirsh, Proc. Natl. Acad. Sci. **102**, 9541-46 and 14475 (2005).
39. V. Mustonen and M. Lässig, to be published.
40. A.M. Moses, D.Y. Chiang, A.P. Pollard, N.I. Iyer, and M.B. Eisen, Genome Biology **5**, R:98 (2004).
41. K. Robison, A.M. McGuire, and G.M. Church, J. Mol. Biol. **284**, 241-254 (1998).
42. S. MacArthur and J. Brookfield, Mol. Biol. Evol. **21**, 1064-73 (2004).
43. D. Arnosti, Ann. Review Entymology **48**, 579-602 (2003).
44. N. Buchler, U. Gerland, and T.Hwa, Proc. Natl. Acad. Sci. **100**, 5136-41 (2003).
45. M. Ludwig, C. Bergman, N. Patel, and M. Kreitman, Nature **403**, 564-67 (2000).
46. M. Ludwig, N. Patel, and M. Kreitman, Development **125**, 949-58 (1998).
47. A. McGregor, P. Shaw, J. Hancock, D. Bopp, M. Hediger, N. Wratten, and G. Dover, Evolution and Development **3**, 397-407 (2001).
48. E. Dermitzakis, C. Bergman, and A. Clark, Mol. Biol. Evol. **20**, 703-14 (2002).
49. J. Scemama, M. Hunter, J. McCallum, V. Prince, and E. Stellwag, J. Exp. Zool. **294**, 285–299 (2002).
50. J. Costas, F. Casares, and J. Vieira, Gene **310**, 215-20 (2003).
51. A. Wagner, Genome Biology **3**, 1012, 1-3 (2002).
52. M. Lynch and J.S. Conery, J. Struct. Funct. Genomics **3**, 35-44 (2003).
53. M. Lynch and J.S. Conery, Science **302**, 1401-04 (2003).
54. E. Davidson, Current Opinion in Genetics & Development **9**, 530-41 (1999).
55. A.P Gasch, A.M. Moses, D.Y. Chiang, H.B. Fraser, M. Berardini, and M.B. Eisen, PLoS Biol. **2**, e398 (2004).